

Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-04-05 15:53:54

PAGE 1

REFERENCE NO: 269

This contribution was submitted to the National Science Foundation as part of the NSF CI 2030 planning activity through an NSF Request for Information, https://www.nsf.gov/publications/pub_summ.jsp?ods_key=nsf17031. Consideration of this contribution in NSF's planning process and any NSF-provided public accessibility of this document does not constitute approval of the content by NSF or the US Government. The opinions and views expressed herein are those of the author(s) and do not necessarily reflect those of the NSF or the US Government. The content of this submission is protected by the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (<https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>).

Author Names & Affiliations

- Jay Walsh - Northwestern University
- Sean Reynolds - Northwestern University

Contact Email Address (for NSF use only)

(Hidden)

Research Domain, discipline, and sub-discipline

Vice President for Research / VP IT and CIO

Title of Submission

Adjusting to the rapid growth of digital research

Abstract (maximum ~200 words).

There continues to be rapid growth of digital research across all fields of research. In emerging areas, such as the social and economic sciences and biological and life sciences, adoption of research cyberinfrastructure has been inhibited by the complexity of the tools, lack of experience leveraging traditional forms of computing, and availability of new tools. As research drives new requirements, being able to design and provision resources from traditional environments has become more challenging. Furthermore, the changing information security landscape continues to increase risk of exposure, making secure environments even more critical.

We believe the top priorities are: 1) continued support and investment in foundational research cyberinfrastructure, 2) focus on developing approaches and practices to support sensitive and regulated data, 3) developing foundational skills and providing support for advanced software development in all areas of computational science and digital research, 4) develop retention and curation models and services for research data at a national scale, 5) improving access to research cyberinfrastructure across all fields of digital research, 6) improving access, guidance and frameworks for the adoption of compliant and scalable cloud computing, and 7) build pathways for developing and funding support practitioners to aid researchers in adoption of technology solutions.

Question 1 Research Challenge(s) (maximum ~1200 words): Describe current or emerging science or engineering research challenge(s), providing context in terms of recent research activities and standing questions in the field.

* Continued support and investment in foundational research cyberinfrastructure

Cyberinfrastructure requirements for research continue to evolve. As researchers collect increasing amounts of data, scalable and affordable data management solutions, including data storage, data transfer methods, and data archival solutions are necessary. This

Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-04-05 15:53:54

PAGE 2

REFERENCE NO: 269

growth is not just occurring in traditional areas of computational research, such as engineering and lab sciences – there have been steady increases in the use of computing and data analysis in fields such as the social sciences and medicine. In addition, researchers in non-traditional fields are rapidly generating and using larger amounts of data. However, scaling workflows from local resources to a national center and gaining rapid access to computing remain difficult challenges that must be addressed. Often times, simply moving codes from one platform to another can be a challenge. As agencies, such as DOE and NSF, continue to drive computing at higher scales, the need for nationally strategic investments that encourage further local investment and enable integration into the national CI ecosystem must continue.

* Focus on developing approaches and practices to support sensitive and regulated data

The nature of the information security landscape is one of continuing evolution and escalation, with more impactful threats, new technologies, and larger impacts on our technology infrastructure.

Contractually and legally restricted data have also shown significant growth and require secure computing, storage, and data movement solutions beyond traditional options. These data normally fall under restrictions, such as HIPAA, FISMA, proprietary data requirements, etc. and affect researchers from a variety of fields.

Additional regulations, such as those covering controlled unclassified information as posed by NIST 800-171, are pressing universities and the research community to rapidly understand and comply with new cyberinfrastructure and end-user practices. However, as the amount of data increases and regulations around cyberinfrastructure security constantly evolve, this task has become difficult for university IT staff to provide affordable environments and workflows that do not inhibit research while still complying with data restrictions. As regulatory requirements continue to evolve, guidance and best practices from the NSF (and other funding agencies), along with programs needed to fund the ongoing compliance and audit activities, become necessary to sustain local research activities.

These challenges will continue to exist across both on premise and off premise (regional, national, and public cloud) research environments. Secure solutions must be available and funded to ensure researchers can continue their research in secure environments.

* Developing foundational skills and support for advanced software development in all areas of computational science and digital research
Growth in research computing also requires the ongoing development of skills for students and faculty. Technology is changing rapidly, and researchers often lack the resources to navigate the options and acquire the practical computational skills now necessary. Fields that are new to computation do not have the background to utilize HPC resources, transfer large amount of data, or write programs and scripts for simulation, processing, or analysis. Fields traditional to HPC are faced with developing codes of increasing complexity and scale. Research and university staff positions must be funded to continue developing these skills, to provide ongoing training to students and faculty, and advance the development of HPC applications.

Researchers that are new to computing require introductory training and skills development in interfacing with the cyberinfrastructure, computer programming, data management, and reproducibility. Training is required for various skill levels, and through different venues such as bootcamps, incorporation to curriculum, smaller workshops, one-on-one sessions, and virtual training. Providing these methods address the diverse needs across research fields, from engineering to digital humanities. We have found that in-person training is often preferred. However, faculty and students are increasingly interested in on demand training for a variety of reasons, making virtual training an important option.

In many research areas, computational workflows and software are still in their infancy and have not yet leveraged parallel computing throughout the field. In fields that are already using high-performance computing, such as materials science, mechanical engineering, and climate science, software development must continue to increase efficiency as data grows and computational capacity increases. This is also true in statistical software, which tend to use large amounts of memory, but cannot utilize memory across nodes in a high-performance computing environment.

Furthermore, funding for computational resources and development of tools is less available in the social and data sciences than in other areas. As the need for computational and storage resources increases in fields such as economics, political science, and psychology, we must find ways to support these efforts.

* Development of retention and curation models and services for research data at a national scale

The scale and variety of research data sets continues to increase over time. Where large projects were once measured in megabytes to gigabytes, multi-terabyte to multi-hundred-terabyte projects are quickly becoming the norm in some areas of research. Additionally, the fundamental changes occurring in areas such as the social and data sciences are resulting in the emergence of new technologies,

Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-04-05 15:53:54

PAGE 3

REFERENCE NO: 269

analytical approaches and ever growing challenges in assembling, managing, and analyzing multiple, diverse sets of data.

Additional requirements around data retention for grant compliance introduces new challenges for universities. Archival solutions are often needed to save data for the length of the grant and beyond. Moving data to a central location is ideal for data loss prevention and data security. Therefore, funding for the creation of archival solutions is essential, with the flexibility to build data archive storage on and off premise. This flexibility will allow institutions to create solutions that are best fit for their research use cases.

We must build upon existing guidelines and regulatory requirements with new education, expertise, and curation services that help our researchers manage their data both during and after execution of a research project. The NSF should continue to identify and fund programs that raise awareness of the importance of data management, seek to build fundamental data management skills in all researchers, and improve sharing of tools and methods across communities of research. In parallel, funding for technical/support expertise continues to be necessary as researchers seek help in identifying communities of practice, solutions, and methods to meet their research data management challenges.

Lastly, to comply with data use agreements and grants, we also see the need to provide the appropriate technology and data storage infrastructure. Affordable and easy to use data management services are essential for research reproducibility and grant compliance when required to keep the data for a certain length of time. New approaches that scale from local to regional and national facilities are required to ensure adoption of appropriate tools and services in financially efficient ways. Some funding programs allow for data management costs beyond the end of the grant. However, keeping every image taken, every data set generated, or the results of every simulation is unsustainable. This can lead to the data statements required for every proposal to becoming diffuse. Analysis of what is truly sustainable in this area (what data must be kept, how should that data be protected and archived, and what can be thrown away) is necessary in order to build solutions that can be effectively adopted at incremental cost.

Question 2 Cyberinfrastructure Needed to Address the Research Challenge(s) (maximum ~1200 words): Describe any limitations or absence of existing cyberinfrastructure, and/or specific technical advancements in cyberinfrastructure (e.g. advanced computing, data infrastructure, software infrastructure, applications, networking, cybersecurity), that must be addressed to accomplish the identified research challenge(s).

* Improving access to research cyberinfrastructure across all fields of digital research

Identity and Collaboration: The rapid growth of tools and technologies are being met with increased levels of collaboration and globalization within research. Online services offered are greater in both number and complexity, and more parts of the daily activities on the University community are predicated on easy access to these services. More and more services are being offered off-campus from the cloud, and maintaining control over identity-related information/attributes is increasingly important in the face of ever-growing security threats and the increased regulations they engender. As services become available outside on premise systems and through web services, they are expected to be available for self-service in real-time, and integrated with one another. As a result, research is growing and becoming more complex as:

- universities enter into more partnership and affiliate agreements with external institutions;
- the scope of our academic institutions becomes increasingly distributed geographically (e.g., other cities in the U.S., partnerships with international schools, and the global spread of research engagements and student learning experiences);
- collaboration with people outside the traditional boundaries of academic research becomes “the new normal” (e.g., fellow researchers at other institutions, consortiums of universities offering courses, peer administrators at other institutions, practitioners outside the University, community engagement);

As these trends continue, identity and access management systems will need to handle a wider variety of situations, offer more options, depend less on physical proximity, and be flexible enough to be deployed quickly and effectively “at scale.” This will ensure they do not become the bottleneck for the deployment of new services, nor an increased risk for compromise. Establishment of and consistent application of these platforms across research cyberinfrastructure must be a priority, along with a renewed focus on supportive infrastructure scalable at national and international levels.

Cyberinfrastructure Broker: The vast array of ever-growing number of new tools and technology introduce new challenges in effective selection, scaling, and adoption of those services. Researchers rely on several types of infrastructure to support their research programs

Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-04-05 15:53:54

PAGE 4

REFERENCE NO: 269

—desktops, servers, local HPC, national facilities, and public/commercial solutions. Each of these has its own unique challenges in adoption into the research workflow. Companies such as ReScale and CycleCloud are attempting to simplify access to computational services by building arbitration layers that enable researchers to effectively select the appropriate computational platform (federally funded, locally funded, and commercially available). Even so, the tools provided are in their infancy. As options continue to grow, further focus on the evolution of Science Gateways (or similar platforms) to more flexibly integrate the expanding set of available tools and technologies is needed.

* Improving access, guidance and frameworks for the adoption of compliant and scalable cloud computing solutions

Across all research fields, access to on-demand computational and data resources are increasing. For some research groups, data is collected quarterly or less per year, requiring bursts of computational needs. For others, their computing and data requirements are beginning to exceed their desktop and laptop capabilities, but large-scale solutions such as clusters and high performance computing environments are overkill. In other cases, data use agreements require stand-alone computing environment, often times these environments cannot be available via the internet. In the past, data was small enough to be performed on individual desktop machines. As data grows, this traditional model breaks down; we are seeing the need for more computational and elastic storage resources for data requiring secure environments. Additionally, cloud-based solutions provide pathways for researchers to easily share workflows and pipelines across institutions. This is helpful in fostering and enabling cross-institutional collaborations, which are often difficult on university clusters and storage due to authentication restrictions.

Areas of further exploration should include:

- leveraging cloud when demand outstrips capacity at a local level,
- testing of new technology solutions such as accelerators and co-processors prior to production adoption,
- infrequent, burst needs for computing and data storage resources that fall short for defendable on premise capital investments in hardware and infrastructure,
- provision of secure, scalable and compliant computing environments to meet regulatory requirements

Lastly, a convergence around federal and campus funding approaches for elastic computing is necessary. Funding streams continue to encourage on premise capital expenditures resulting in physical machines as compared to shared services. This consistently leads to confusion not just with expenditures on central on premise HPC systems, but also expenditures associated with external/cloud services. Additionally, more specific to cloud computing, the transition from capital to operating expenditures funding means that gaps in research funding (either locally or nationally driven) must be accounted for by the researcher so that access to computing and storage services can be sustained to drive their research.

An organized approach for use of cloud computing that emphasizes standards, repeatability, and best practices in security and compliance is needed nationally. Guidance of how funded research can use grant dollars to purchase cloud resources remains vague at best, and must be addressed. A sharp focus on the development of guidance and frameworks for use of public cloud resources is well past due.

* Build pathways for developing and funding support practitioners to aid researchers in adoption of technology solutions.

Cyberinfrastructure is not only comprised of wires, switches and hardware; it includes people. Regardless of how resources for compute and storage are provisioned (HPC, cloud, etc.) researchers are increasingly depending upon research consultants/CI engineers/ Research Software Engineers/et al. to understand what tools will work best for them and how to make those tools fit into their research workflows.

The career path for these individuals needs attention. Support for programs that establish and maintain career paths has already become necessary. Development of a thriving professional community (possibly rooted through the growth of ACI-REF) is foundational to establishing this role as viable career option. Additionally, improved models are needed for employing these individuals that does not lead to gaps in their funding. A new model that combines agency and university funding, and is flexible across individual grants, has become necessary. These challenges are even more pronounced in areas such as the social sciences and humanities where funding has historically not accounted for these needs.

Question 3 Other considerations (maximum ~1200 words, optional): Any other relevant aspects, such as organization, process, learning and workforce development, access, and sustainability, that need to be addressed; or any other issues that NSF should consider.

Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-04-05 15:53:54

PAGE 5

REFERENCE NO: 269

To address the increasing security requirements and the use of new cyberinfrastructure technologies, guidance around architecting new services will be necessary. As we look to use the cloud and additional shared resources, we will need to have CI practitioners who increasingly understand both the technology and security requirements of research projects. These new roles will also be needed for creating guidelines and policies for researchers using the ecosystem of computing and data services.

Software needs to be further developed to take advantage of computing and memory that is available on modern technologies. As cyberinfrastructure improves, software commonly used in engineering simulations, statistics analysis, and bioinformatics pipelines also need continuous development. Currently, NCAR has shown performances of climate modeling codes on modern supercomputers operate at 1% efficiency. This is entirely too low to effectively leverage investments made by NSF and other agencies in leadership class computing systems and may impact our ability to address challenges facing us at national and global levels. As another example, most genomics pipelines do not leverage the full power of high-performance computing clusters, providing opportunity for improvements that positively impact both research and clinical care. To increase research productivity, funding must continue to be provided to hire capable students, postdocs, and staff that provide ongoing software support and development at scale.

In fields such as political science, psychology, business, economics, and bioinformatics, programming and computing skills are often not developed within the traditional curriculum. In addition, we find, at times, graduate students in engineering and lab sciences are also unprepared to leverage computing and programming environments when they arrive. This has led to a pressing need for skill development to clean and analyze data sets, use shared Linux computing facilities, basic programming, and data visualization. Currently, training is provided at Northwestern through a variety of departments, schools, and units. In addition, Northwestern has recently launched an online programming bootcamp to learn foundational data-science related languages: R and Python.

Even with these programs in place, demand for foundational skills continues to outstrip our capacity to support and educate our community. Unfortunately, the staff time to provide on-site workshops is not scalable. Funding for skills development workshops for computing and visualization, as these skills are increasingly needed in various research fields. In addition, online training provides students and faculty with an opportunity to learn skills on their own time. Trainings that are currently provided, such as those hosted by XSEDE, will continue to be essential. Longer and more in-depth programming workshops, that focus on specific programming languages, like Python, should also be provided at a national level. In addition, funding for travel to these trainings will be necessary for students and faculty.

As the research landscape changes, we must adjust with it. Building skills in our research community and in our university staff is essential to continue to perform world class research that continues to push new boundaries across engineering, lab sciences, social sciences, business, and medicine. Training opportunities and additional staff will be needed to support research on the horizon. Workforce development is essential to update security controls, policies, and cyberinfrastructure on campuses nationwide. We strongly encourage the NSF to increase funding for training opportunities, while offering additional nation-wide training for computational research skills.

Consent Statement

- "I hereby agree to give the National Science Foundation (NSF) the right to use this information for the purposes stated above and to display it on a publically available website, consistent with the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (<https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>)."
-